Computational Aesthetics in Graphics, Visualization, and Imaging (2009)
O. Deussen and P. Hall (Editors)

Movie Posters from Video by Example

S. Brooks [1]

[1]Faculty of Computer Science, Dalhousie University, Canada

Abstract

We present the first method for the generation of posters directly from video, borrowing the layout and composition from existing poster exemplars as a guide. Our system analyzes a given poster by determining face locations and poses, detecting the title and computing an analysis of the remaining background image. Processing of the video proceeds by locating major cast members and a suitable background frame. A new poster is then seamlessly constructed from scratch with faces, title and background appropriately sized and positioned as in the example. This work has broad application and potential for widespread use given the increasing importance of Creative Commons amateur film making, as well as internet and personal video.

Categories and Subject Descriptors (according to ACM CCS): I.3.3 [Computer Graphics]: Applications I.3.3 [Computer Graphics]: Picture/Image Generation

1. Introduction

Recently personal video creation has surged due to low-cost and miniaturized video cameras. Video is now made with cell phones, digital cameras, digital video cameras and webcams. This trend has been bolstered by the rise of broadband which has itself spawned a huge interest in internet delivered personal video. Examples of which are Google videos, YouTube videos and amateur films created under the Creative Commons Licence.

With a vast selection of video on offer over the internet, the selection of video becomes key. This is now performed by categorical browsing or the browsing of video selections returned by query, with each video represented by an extracted still image. Our aim is to provide an alternative to this by generating a poster which represents the video's main characters while stylistically borrowing from existing movie posters. An example is shown in Figure 1. Note that original posters have been obscured with NPR filters. In addition, we believe this work could find broader application including browsing one's own personal video collection and recorded TV shows on DVR, as well as automatically generating personal DVD covers, DVD menu art and DVD disc art.

In most cases, generating posters by example requires face and pose detection, text detection, as well as image analysis and compositing. Although many more deserve mention, due to the extensive list of related work, we now review only those papers most related to the present work.

2. Related Work

As face detection constitutes a crucial aspect of the current work, we begin noting a few key recent papers [Row99, SK00]. In particular, the work by Viola and Jones [VJ01] achieves high accuracy in real-time using a cascade of weak detectors. Schneiderman [Sch04] later improved the accuracy of cascaded classification with a feature-centric approach.

In addition to detecting faces, we are also concerned with determining facial pose and expression where possible. Early work in facial recognition by Turk and Pentland [TP91] introduced the eigenface approach. Though innovative, was not robust to changes in pose and expression. In the case of Active Appearance Models (AAM) [6], facial shape and texture properties are combined in an iterative search process to determine facial pose. But as we will later discuss, the major drawback of this approach is that it can only determine local minima and is highly sensitive to initial search conditions.

The most closely related work is that by Blanz et al. which is able to construct 3D facial models from images [BV99], re-animate faces [BBVP03] and replace faces [BSVS04]. However, they do not locate faces in images, instead relying

on manual positioning or feature point labeling. Regarding face replacement, one might argue that results are not completely realistic and the technique requires semi-automatic hair masking. More generally, these systems do not construct new posters from video which involves additional text analysis and video summarization, nor do they consider ensembles of faces. Also, in our case no content from the exemplar poster is directly used.

Video analysis is also related to the current work. For example, video abstraction distills a longer video into images or shorter videos [LLZ*01]. Some systems have also attempted to make use of face detection for video abstraction [LPE97], cast listings and video indexing [EWIR01].

Further systems have been published for text detection in both video and still images [WJC02]. Smith and Kanade [SK95] present an intra-frame method to detect text in video, while Lienhart et al. [LPE97] use text detection for the purpose of facilitating video abstraction.

Finally, poster generation also relates to recent work on image compositing [PGB03], [JSTS06] and collages [RBHB06], [JBS*06]. Rother et al. [RBHB06] uses face detection to ensure that faces are kept whole within the generated collages and Johnson et al. [JBS*06] generate composite images from positioned words on a canvas, such as "boat" and "sand".

3. Exemplar Poster Analysis

The first stage of the process involves the decomposition of the exemplar poster into face, title and background regions.

3.1. Face Location and Pose Detection

Our criteria for face detection differ from those of typical face detection research:

1. we do not require real-time performance,
2. we are more concerned with the group of faces rather than just individual faces, and
3. as it is not a life or death application, we aim for approximately correct overall results by minimizing the occurrence of truly degenerate cases, rather than attempting to maximize the percentage of perfect matches.

We therefore proceed in a different fashion than prior work. First we detect the face locations with a cascaded detector. We then attempt to determine face poses with a set of AAMs. Finally, we minimize the chance of selecting a degenerate pose from the outputs of the AAMs by considering *pose ensembles*.

When properly initialized, AAMs are adept at computing facial poses; however, we cannot directly use AAMs [6] without first locating faces with a separate method since AAMs are highly sensitive to initial conditions. To locate faces we utilize the existing work of Schneiderman [Sch04].

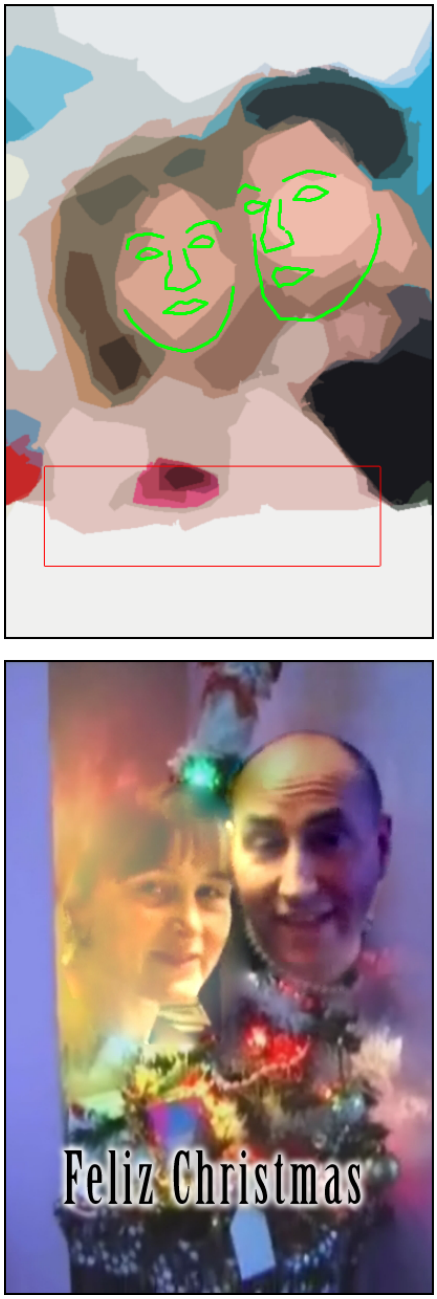

Figure 1: *New poster (bottom) generated by example (top).*

This is similar to previous cascaded detectors but significantly improves results by efficiently sharing feature values among overlapping detection windows. Note that we discard very small and very low contrast faces. Example results of this initial stage are shown as the smallest blue squares for each face, as shown in Figure 2, top.

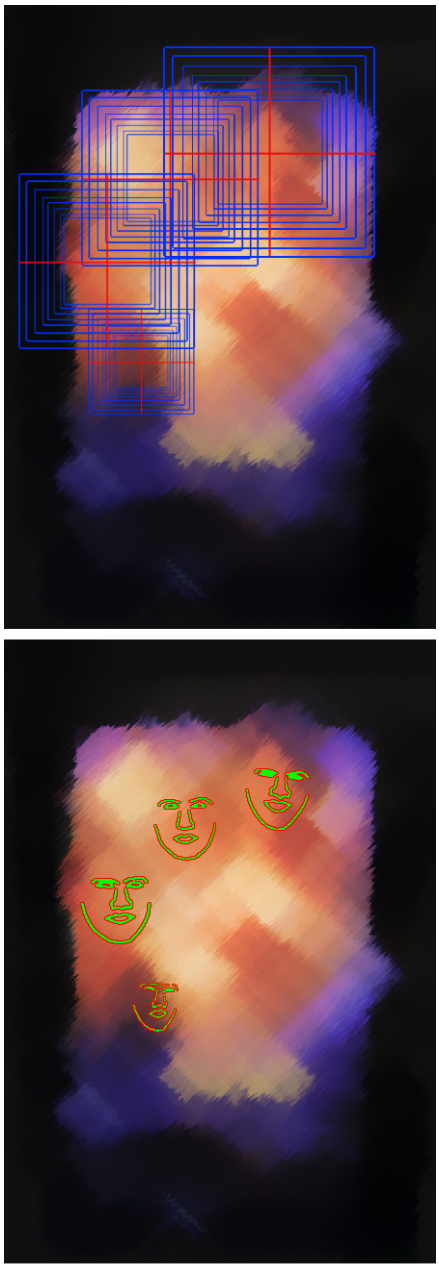

Figure 2: *Face locations (top) and poses (bottom) detected.*

Once faces are located, we determine each pose with a set of AAMs, which are statistical models of facial shape and texture. AAMs combine a model of shape variation with a model of shape normalized texture variation, both learned from a training set of labeled images. In our case, labeling consists of 60 landmark points per face. The shape of a given face is approximated by:

$$x = \bar{x} + P_s b_s$$

where $\bar{x}$ is the mean shape, P_s are orthogonal modes of shape variation and b_s are face shape parameters. Likewise, facial texture is modeled as:

$$g = \bar{g} + P_g b_g$$

where $\bar{g}$ is the shape and lighting normalized mean texture, P_g are the orthogonal modes of texture variation and b_g are texture parameters for the face. Modeling a particular face image requires a multi-resolution optimization of all model parameters.

Faces in posters are front facing, rotated left or rotated right, since significant downward or upward tilts are rare. Minor tilts are not an issue. We therefore construct 6 distinct AAMs (front, left and right for both male and female subjects), each learned from 40 hand marked 2D images. The median face for each face model is shown in Figure 3. Although this covers the pose range, the estimated face size from the cascaded detector is not always accurate, and our experiments show AAMs to be highly sensitive to initial size estimates. We therefore run the 6 AAMs at multiple initial sizes (shown as 8 blue boxes for each face in Figure 2, top).

Another limitation of AAMs is the lack of treatment of background clutter, since the training faces are assumed to reside over blank backgrounds. We cannot make that assumption in posters and our empirical experiments show that AAMs can latch onto background detail. And so, we extend AAMs to mitigate the presence of background content. This is done by first constructing an outer mask as a dilated union of all 6 median face shapes, $\bar{x}$:

$$m_1 = \bigcup_{i=1}^{6} \lceil \bar{x}_i / 255 \rceil \oplus A$$

where A is a disk of radius 10 pixels. A second inner mask, m_2, is constructed in a similar fashion, but with an erosion operation. The two masks are used as a basis for a GrabCut [RKB04] between m_1 and m_2. We then apply a Gaussian blur of radius 15 to the background. The blur is linearly ramped inwards to the face over a distance of 20 pixels in order to avoid introducing new discontinuities. These values are relative to a face of 300 x 300 pixels. The resulting image is subsequently fed into each AAM for pose detection.

This generates good results for most faces, but it is still possible that a good match will not be found. In addition, we need to gracefully handle false positives from the cascaded detector. Distinguishing good matches from bad is not difficult as error rates for mismatches are 5 times higher on average. For mismatches, we guess a reasonable pose with respect to other faces in a *pose ensemble*. Given the size and location of all faces in the exemplar poster, we predict poses

Figure 3: *Six active appearance model means.*

for poorly matched faces using pose ensembles in an existing database of posters. This is done by finding the three most similar posters with the same number of faces. The difference between posters P and Q is computed as:

$$d_{pq} = min_{N(k),\,M(k)} \left(\sum_{i,j} L_2 \left(v_i^p, v_j^q \right) \times \frac{max\left(s_i^p, s_j^q \right)}{min\left(s_i^p, s_j^q \right)} \right)$$

where N, M are permutations of $(1..k)$ faces, $i \in N(k)$, $j \in M(k)$, v_i are face centers and s_i are face sizes. We then use the corresponding pose in each of the three posters to initialize new AAM searches. Searches are performed for male and female AAMs with the closest mean face shape, at all scales. If this produces a good match, it is used; otherwise, we directly use the corresponding pose of the most similar poster as a final default. This default may not be 100% correct but it also will not usually be severely anomalous.

3.2. Text Detection

Our approach to text detection is conservative, since there can be a wide variety of objects that a detector may falsely classify as text. We therefore restrict our search to finding only the title text of the exemplar poster and, once again, we aim for plausible results, rather than maximizing the percent of perfect matches.

In addition to eliminating facial areas, we can also make use of approximate regularities in all posters to estimate where the title may lie. For this we return to the concept of ensemble, this time including both faces and title text. Using the same three matches in the database, we search for the largest text region within the dilated union of their three title locations (Figure 5, middle). The union is dilated by an elliptical structuring element with a width radius of 50 pixels and a height radius of 25. We also assume that the detected text region is at least 50% of the area of the average database match, since small text is unlikely to be the title. For the text detection itself we use [WJC02] which is based on accumulated gradients and morphological processing, though any suitable text detector might serve. If the title is not detected given these constraints, we directly use the corresponding title region from the best match.

3.3. Background Analysis

Background analysis must be reasonably fast and approximate since we will use the same analysis on frames of the input video. We also restrict the analysis to color properties since color harmony is an important aspect of posters, and we have no reason to believe the background content of the exemplar poster bears any relation to the video content. For this, a feature vector, C, is calculated over the original poster with dilated facial and title regions removed. Dilation is performed with a disc of size 10. The color feature vector is computed as a color histogram with a bin width of 20, in the perceptually-based L*a*b* space. This yields 5 bins in the L* dimension and 10 bins in each of the a* and b* dimensions. This totals 500 bins, though only 218 bins fall within the gamut corresponding to $0 < (R,G,B) < 1$.

4. Video Analysis

The second stage requires an analysis of the input video to detect faces and determine suitable background images.

4.1. Face and Pose Detection in Video

Detecting faces in video is less difficult since we do not aim to detect all characters in the video, just the main ones; nor do we need to detect all instances of each main character. We begin by applying a cascaded face detector on all frames, f_i, detecting as many faces, v_{ij}, as the algorithm produces, where i is the frame number, and j is the number of the face within that frame. We then cluster these into character classes, V_k, using a neural network approach [LPE97]. That is, each v_{ij} is put into a V_k. We retain the top 5 character classes based on the number of frame members. We only retain the top 5 characters since we do not want well-matched minor characters. Then for each face, p_n, in the poster we find a set of 10 best pose matches in v_{ij} such that there is at least one match from each class V_k. To determine the best matching poses we apply the corresponding AAM that detected the current face's pose in the original poster. For example, if face p_n was detected with the front facing female AAM, then we attempt to locate up to 10 best matches in frames, v_{ij}, using that same AAM. The reason for retaining 10 best matches, rather than a single best match, is given in section 5.

From the 10 matches per poster face, p_n, we assign each p_n a best match. When assigning best matches, we minimize the total pose error over all p while requiring that each poster face is assigned a match from a unique face class, V_k.

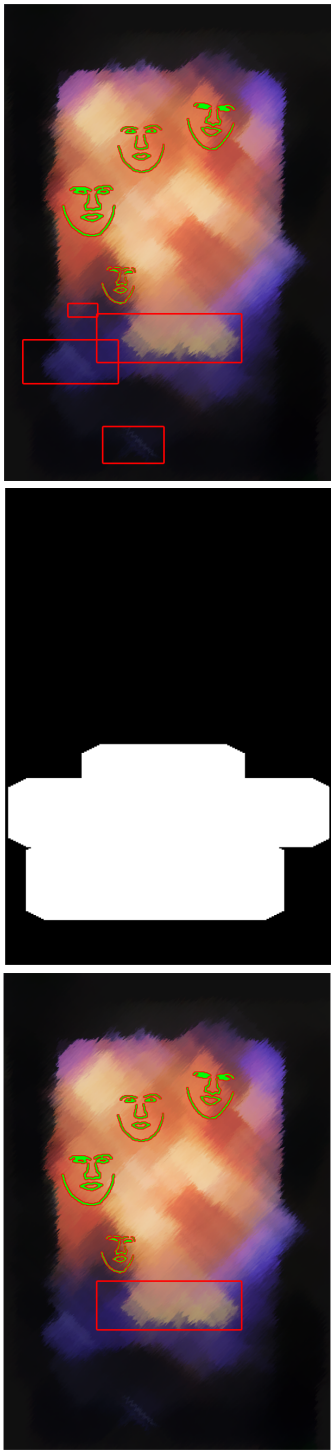

Figure 4: *All 'text' detected in non-face areas (top), the dilated union of matched title regions (middle) and the final title area (bottom).*

4.2. Finding Background Images

Suitable background images also need to be selected from the video. Ideally we would like to choose an interesting frame that matches the existing color scheme of the original poster.

The search is conducted on all frames after eliminating those containing faces and text, and this remaining set of frames is further broken down into 'special event' and regular frames [LPE97]. Note that special events are currently limited to gun fire and explosions as determined from audio analysis, but could be extended to other audio-detectable actions.

If special events are detected we select up to five frames with the closest matching color histograms, by computing the same color feature vector as in described in section 3.3. These frames are chosen using the following distance metric between poster image m and frame n:

$$d^2_{hist}(C_m, C_n) = (C_m - C_n)^T A (C_m - C_n)$$

where C_m,C_n are the histograms for poster image m and frame n as described in section 3.3, $A = (a_{ij})$ is a symmetric matrix of weights between 0 and 1 representing the similarity between bins i and j based on the distance between bin centers. Neighboring bins are given a weight of 0.75, bins that are 2 and 3 units away are assigned weights of 0.5 and 0.25. All other weightings are set to 0. This weighting makes the distance calculation more robust to minor variation.

If present, the top matching special event frame is used as the default background image. A further 5 to 10 top matching frames are then selected from the non-special event frames. Exactly how many are needed depends on how many special frames were found since a combined total of 10 potential frames are taken.

5. User Overrides

We offer the user simple controls to override aspects of the final result where they deem it desirable. The principle is to first provide the user with a best guess, then allow the user to replace individual poster elements using a small set of provided alternatives. For each face and for the background image, the user is given the option of replacing the best match with any of the remaining 9 matches. If this proves insufficient, additional sets of override choices can be presented to the user in subsequent screens. Replacing the best matching face with a near-best match requires two mouse clicks which indicate the swap-pair.

6. Compositing the New Poster

To form the final poster we first insert the background image (cropped to center) taken from the video into a blank image. Then we sequentially insert each face. Because determining

how much 'body' is below a face is a very difficult problem in general, we conservatively restrict the inserted face region to the head area in all cases. Starting with an outer rectangular region around the face, the region is further optimized using an iterative graph cut [RKB04]. Each face's region is then dilated by 10 pixels and sequentially integrated into the poster using gradient domain pasting [PGB03]. Finally, the title is inserted using a default font. The title of the new poster is taken directly from the filename of the input video.

7. Results

In addition to Figure 1, further results are shown in Figures 5- 6. We have used the same exemplar poster for multiple video sources in Figure 6 in order to show the flexibility of poster/video combinations. There are a few cases in which overrides were required that are worth noting. The "Feliz Christmas" example in Figure 1 required a user-override of the background image to better capture the theme of the video, as did the "Balloon Ballistic" example in Figure 6.

Figure 5 shows a case in which the face pose detector fails. The middle face's pose in the original poster was not successfully detected, probably due to harsh lighting conditions. However, the pose ensemble predicted a conservative pose tilted slightly to the right. Although this guessed pose might not have been accurate in all cases, it is also not likely to be overly degenerate.

The source locations of all the videos used are noted in the appendix. All examples are generated from Creative Commons amateur documentaries.

8. Limitations

Currently our approach to face detection can suffer under the presence of harsh lighting conditions. Determining such conditions and treating them as special cases might prove useful. Another limitation of our method is an inability to detect body areas below each face and a technique for font matching would also be beneficial.

More generally, our work could be considered a proof-of-concept system that would make use of continuing improvements in the fields of face and text recognition. For the present, face detection and recognition remains imperfect but continues to progress. Understanding this, we consciously designed the system to be as fault-tolerant as possible and provide the user with a two-click override system for replacing components from a set of plausible candidates. Most examples that we tested required a modest amount of user guidance, in the form of one or two user-overrides.

However, we also note that the limitations of face detection are not as critical with regards to video processing in this application, since it is not necessary for every face in every frame to be detected in the input videos. We only require a representative image set for key characters in the video. As major characters appear frequently in a typical video, there tend to be a sufficient number of frames with clear poses to extract every character at least once.

9. Conclusions and Future Work

We have presented a method to construct movie posters directly from video, while borrowing stylistic elements from real movie posters. The potential applications of this are broad and there remain significant opportunities to extend our system. The selection of poster exemplars could also be explored. For example, one might attempt to automatically determine the genre of the input video and match it with an appropriate exemplar or standard template in a database. Also, a design gallery approach might be incorporate aspects such as text style, poster genre and decade. Another interesting extension would be the incorporation of video textures to produce dynamic posters. Furthermore, applying standard non-photorealistic filters to faces and/or background images would create more stylized posters.

Other compelling application areas include the generation of post cards and greeting cards directly from video. For example, given a home movie of a family vacation, one might generate personalized post cards.

Lastly, the work could be made more robust if developed as a commercial system by resticting the degree of generality and automation. For example, the system could operate with poster template styles rather than exemplars.

10. Acknowledgements

This work was supported by the Natural Sciences and Engineering Research Council of Canada (NSERC) and the Canada Foundation for Innovation (CFI).

References

[BBVP03] BLANZ V., BASSO C., VETTER T., POGGIO T.: Reanimating faces in images and video. In *EUROGRAPHICS 2003* (Granada, Spain, 2003), vol. 22, pp. 641–650.

[BSVS04] BLANZ V., SCHERBAUM K., VETTER T., SEIDEL H.-P.: Exchanging faces in images. In *EUROGRAPHICS 2004* (Grenoble, France, 2004), Cani M.-P., Slater M., (Eds.), vol. 23 of *Computer Graphics Forum*, Blackwell, pp. 669–676.

[BV99] BLANZ V., VETTER T.: A morphable model for the synthesis of 3d faces. In *SIGGRAPH '99* (New York, NY, USA, 1999), ACM Press/Addison-Wesley Publishing Co., pp. 187–194.

[EWIR01] EICKELER S., WALLHOFF F., IURGEL U., RIGOLL G.: Content-Based Indexing of Images and Video Using Face Detection and Recognition Methods. In *IEEE Int. Conference on Acoustics, Speech, and Signal Processing (ICASSP)* (Salt Lake City , Utah, 2001).

Figure 5: *Generated poster with 3 faces. The middle face's pose in the original poster was not successfully detected. However, the pose ensemble predicted a conservative pose tilted slightly to the right.*

[JBS*06] JOHNSON M., BROSTOW G. J., SHOTTON J., ARANDJELOVIĆ O., KWATRA V., CIPOLLA R.: Semantic photo synthesis. *Computer Graphics Forum (Proc. Eurographics) 25*, 3 (September 2006), 407–413.

[JSTS06] JIA J., SUN J., TANG C.-K., SHUM H.-Y.: Drag-and-drop pasting. *ACM Transactions on Graphics (SIGGRAPH)* (2006).

[LLZ*01] LI Y., LI Y., ZHANG T., ZHANG T., TRETTER D., TRETTER D.: *An overview of video abstraction techniques*. Tech. rep., Report HPL-2001-191, HP Laboratories, 2001.

[LPE97] LIENHART R., PFEIFFER S., EFFELSBERG W.: Video abstracting. *Commun. ACM 40*, 12 (1997), 54–62.

[PGB03] PÉREZ P., GANGNET M., BLAKE A.: Poisson image editing. *ACM Trans. Graph. 22*, 3 (2003), 313–318.

[RBHB06] ROTHER C., BORDEAUX L., HAMADI Y., BLAKE A.: Autocollage. *ACM Trans. Graph. 25*, 3 (2006), 847–852.

[RKB04] ROTHER C., KOLMOGOROV V., BLAKE A.: "grabcut": interactive foreground extraction using iterated graph cuts. In *SIGGRAPH '04: ACM SIGGRAPH 2004 Papers* (New York, NY, USA, 2004), ACM, pp. 309–314.

[Row99] ROWLEY H. A.: *Neural Network-based Human Face Detection*. PhD thesis, Carnegie Mellon University, 1999.

[Sch04] SCHNEIDERMAN H.: Feature-centric evaluation for efficient cascaded object detection. *Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on 2* (June-2 July 2004), II–29–II–36 Vol.2.

[SK95] SMITH M. A., KANADE T.: *Video skimming for quick browsing based on audio and image characterization*. Tech. rep., 1995.

[SK00] SCHNEIDERMAN H., KANADE T.: A statistical method for 3d object detection applied to faces and cars. *Computer Vision and Pattern Recognition, 2000. Proceedings. IEEE Conference on 1* (2000), 746–751 vol.1.

[TP91] TURK M., PENTLAND A.: Eigenfaces for recognition. *J. Cognitive Neuroscience 3*, 1 (1991), 71–86.

[VJ01] VIOLA P., JONES M.: Rapid object detection using a boosted cascade of simple features. *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on 1* (2001), I–511–I–518 vol.1.

[WJC02] WOLF C., JOLION J., CHASSAING F.: Text localization, enhancement and binarization in multimedia documents. In *International Conference on Pattern Recognition (ICPR) 2002* (2002), pp. 1037–1040.

Appendix A: Video Locations

The following is a list of source locations of the videos used to generate the results in this paper:

- "Feliz Christmas": `mms://video.channel4.com/culture/fourdocs/30285/30285_720.wmv`
- "Laugh out Loud": `mms://video.channel4.com/culture/fourdocs/31705/31705_720.wmv`
- "Behind the Flames": `mms://video.channel4.com/culture/fourdocs/1166/1166_720.wmv`
- "Balloon Ballistic": `mms://video.channel4.com/culture/fourdocs/24845/24845_720.wmv`
- "Cannes Home Movie": `mms://video.channel4.com/culture/fourdocs/4561/4561_720.wmv`

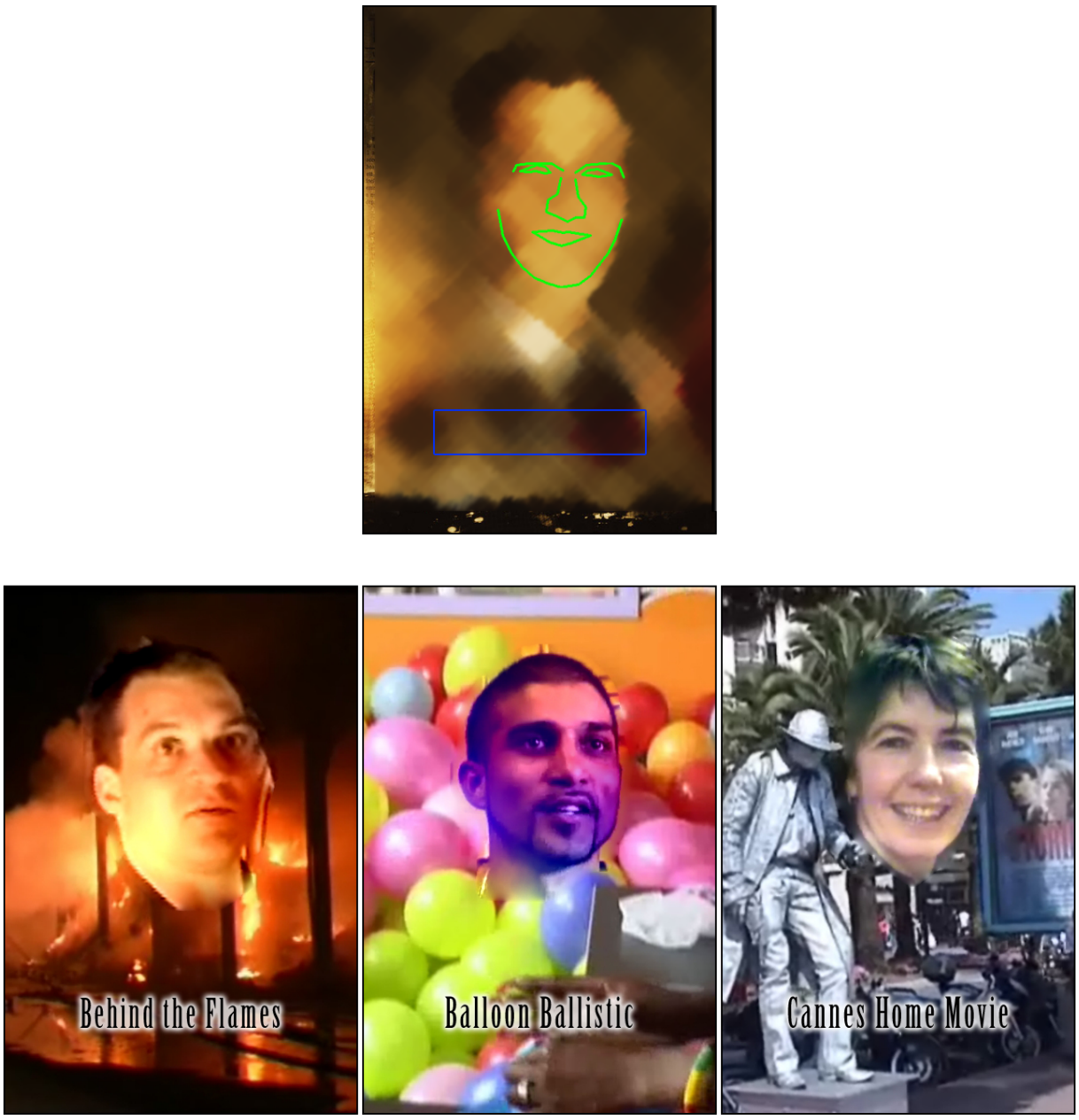

Figure 6: *Set of results using same poster exemplar for multiple video sources. Exemplar poster shown above, with remaining three posters generated from 3 separate video sources.*